

ARVUTITEADUSE INSTITUUT
TARTU ÜLIKOOL

Andmekäevandamise uurimisseminar

Jaak Vilo
(toimetaja)

Tartu 2003

Andmekaevandamise uurimisseminar

Jaak Vilo (toimetaja)

Arvutiteaduse instituut
Tartu ülikool
Liivi 2, 50409 Tartu, Estonia

Raport. Andmekaevandamise uurimisseminar MTAT.03.169

Detsember 2003, vi + 198 lk.

Sisukord

| | |
|--|-----|
| Eessõna | v |
| Assotsiatsioonireeglite leidmine suurtest andmehulkadest <i>Asko Tiidumaa</i> | 1 |
| Otsustuspuudega klassifitseerimine <i>Kristo Käärmann</i> | 16 |
| Induktiivne loogiline programmeerimine <i>Igor Kuzmitšov</i> | 33 |
| MDL — lühima kirjelduse printsiip <i>Meelis Kull</i> | 50 |
| Klasterdamine andmekaevanduses <i>Mihhail Juhkam</i> | 70 |
| Informatsioonikaugus <i>Mart Sõmermaa</i> | 90 |
| Sagedasti esinevate sõnemustrite otsimine ja algoritm Teiresias <i>Ireen Meho</i> | 101 |
| Tõenäosuste leidmine Bayesi võrkudes <i>Sven Laur</i> | 112 |
| Sissejuhatus tugivektor-masinas <i>Hando Tint</i> | 136 |
| Ülevaade EM-algoritmist <i>Jelena Zaitseva</i> | 148 |

| | |
|---|------------|
| Tekstikaevandamine: infoeraldus <i>Jüri Reimand</i> | 161 |
| A Data Clustering Algorithm for Mining Patterns from Event Logs <i>Risto Vaarandi</i> | 178 |
| Kahendklasterdamine <i>Ants Aader</i> | 188 |

Eessõna

Andmekaevandamise (*ingl.k.* Data Mining, DM, või Knowledge Discovery from Databases, KDD) eesmärk on leida andmetest seaduspärasusi, reegleid, trende või muid aspekte mis on kasutajale seni teadmata ja huvitavad. Andmekaevandus on suhteliselt noor uurimisvaldkond mis on saanud mõjutusi eri aladelt nagu statistika, andmebaaside teooria, masinõppimine, algoritmiline arvutiteadus ning loomulikult paljudelt erinevatelt rakendusvaldkondadelt. Üheks andmekaevandust iseloomustavaks omaduseks võrreldes näiteks traditsioonilisema masinõppimise või statistilise analüüsiga on tavaliselt andmete suur maht mis eeldab hästi skaleeruvaid analüüsimetode.

Käesolev kogumik sisaldab erinevate andmekaevandusega seotud valdkondade referatiivseid ülevaateid. Neis käsitletakse traditsioonilisi assotsiatsioonireeglite otsimise meetode (kes ostab õlut ja sinepit korraga ostab suure tõenäosusega ka grillvorsti); masinõppimise meetode ja formalisme — otsustuspuude algoritme, Bayesi võrkude teooriat, tugivektor-masinaid (SVM), induktiivset loogilist programmeerimist; teoreetilisi mõõte õppimistulemuste headuse hindamiseks (lühima kirjelduse printsiip MDL); klasteranalüüsi meetode ning kauguse mõõte; andmete suurima tõepära hinnangu leidmiseks kasutatavat EM-algoritmi; tekstilistest andmetest mustrite otsimist; ning andmekaevanduse rakendusi — tekstide analüüsi, bioinformaatikat, ning arvutiturvet.

Kogumik sisaldab Tartu Ülikooli Arvutiteaduse instituudi Andmekaevanduse uuriseminaris (MTAT.03.169, sügis 2003) osalejate referaate. Seminari idee on saadud Helsingi Ülikooli Arvutiteaduse instituudi sarnasest eksperimentidist (Klusterointimenetelmät-seminaari, 2002, prof. Hannu Toivonen). Seminar matkib tavalist teaduskonverentsi, kus seminaril osalejate ülesandeks oli koostada oma teema kohta referaat (“artikkel”) ning esitada see seminari (“konverents”) korraldajale. Iga referaati hindas omakorda kolm seminaril osalejat (“programmitoimikond”). Saadud arvustuste põhjal sai veel parandada oma referaati misjärel see edastati korraldajatele lõplikule kujule viimiseks. Seminar lõppeb kahepäevase konverentsiga kus iga osaleja esitab oma teema ka suulise ettekandena.

Kuna tegemist on Eestis suhteliselt uue teadusvaldkonnaga, sest andmekaevandust ei ole Tartu Ülikoolis õpetatud, on eksperiment seda julgem. Kuiigi andmekaevandamise teema on enamusele seminaris osalejatele uus, näitasid seminari osalejad üles väga suurt põhjalikkust teemade käsitlemisel. Nii mõnigi referaat on oma sisult ja pikkuselt kaugelt üle seminari ametlikust 3AP programmi tasemest ning vastavad pigem süvaõppe tasemel uurimisteedadega. Juba artiklite esimesed versioonid olid kõrgetasemelised ning kaas-

laste soovitusel ja kriitika aitas neid veelgi parandada.

Korraldajana saadud kogemuste põhjal julgen väita, et seminaris osalemine on osutunud kõikidele osalejatele kasulikuks ennekõike igaühe enda õpingute jaoks, eriti kirjutamise-oskuste lihvimise ning teemasse süvenemise seisukohast. Loodetavasti muutuvad sellised seminarid kirjalikult vormistatud referaatidega, olgu siis iganädalase “journal club” või ka konverentsi vormis, traditsiooniliseks osaks arvutiteaduse alast kõrg- ja teadusharidust.

Suur tänu kõikidele seminaris osalejatele! Soovin kõikidele huvitavat ja kasulikku “konverentsi”! Lisaks soovin ka kõigile teistele käesoleva kogumiku lugejatele, et leiaksite siit kogumikust huvitavat infot ning otsest kasu.

Detsember 2003, Tartu.

Jaak Vilo